1

# PROCOLLAGEN ASSEMBLY

The present invention relates to a method of regulating assembly of procollagens and derivatives thereof.

Most cells, whether simple unicellular organisms or cells from human tissue, are surrounded by an intricate network of macromolecules which is known as the extracellular matrix and which is comprised of a variety of proteins and polysaccharides. The major protein component of this matrix is a family of related proteins called the collagens which are thought to constitute approximately 25% of total proteins in mammals. There are at least 20 genetically distinct types of collagen molecule, some of which are known as fibrillar collagens (collagen types I, II, III, V and XI) because they typically form large fibres, known as collagen fibrils, that may be many mircometers long and may be visualised by electron microscopy.

Collagen fibrils are comprised of polymers of collagen molecules and are produced by a process which involves conversion of procollagen to collagen molecules which then assemble to form the polymer. Procollagen consists of a triple stranded helical domain in the centre of the molecule and has non-helical regions at the amino terminal (known as the N-terminal propeptide) and at the carboxy terminal (known as the C-terminal propeptide). The triple stranded helical domain is made up of three polypeptides which are known as α chains. Procollagen is synthesised intracellularly from pro-α chains (α chains with N- and C-terminal propeptide domains) on membrane-bound ribosomes following which the pro-α chains are inserted into the endoplasmic reticulum.

Within the endoplasmic reticulum the pro-α chains are assembled into procollagen molecules. This assembly can be divided into two stages: an initial recognition event between the pro-α chains which determines chain selectively and then a registration event which leads to correct alignment of the triple helix.

2

Procollagen assembly is initiated by association of the C-terminal propeptide domains of each pro-α chain to form the C-terminal propeptide. Assembly of the triple helix domain then proceeds in a C- to N-terminal direction and is completed by formation of the N-terminal propeptide. The mature procollagen molecules are ultimately secreted into the extracellular environment where they are converted into collagen by the action of Procollagen N-Proteinases (which cleave the N-terminal propeptide) and Procollagen C-Proteinases (which cleave the C-terminal propeptide). Once the propeptides have been removed the collagen molecules thus formed are able to aggregate spontaneously to form the collagen fibrils.

Collagens have many uses industrially. For instance, Collagen gels can be formed from collagen fibrils *in vitro* and may be used to support cell attachment. Such gels may be used in cell culture to maintain the phenotype of certain cells, such as chondrocytes explanted from cartilage. Collagen may be also used as a "stuffer" or packing agent surgically and is particularly known to be used in cosmetic surgery, for enlarging the appearance of lips for instance. *In vivo*, collagen is a major component of the extracellular matrix and serves a multitude of purposes. Numerous diseases are known which involve abnormalities in collagen synthesis and regulation. Procollagens and derivatives thereof may be used (or be of potential use) for the treatment of these diseases.

Large quantities of procollagens or derivatives thereof need to be synthesised to meet increasing industrial demand. A convenient means of synthesising procollagens or derivatives thereof is by expression of exogenous pro-α chains in a host cell followed by the assembly of pro-α chains into the procollagen or derivative thereof. For this to occur it is necessary to ensure that any host cell used has the necessary post-translational facilities required to assemble procollagens from pro-α chains. This may be achieved by expression in cells which normally synthesise procollagen. However one problem in such systems is that endogenously expressed

pro-α chains can co-assemble with the exogenously introduced pro-α chains giving rise to undesirable hybrid molecules.

In other circumstances it may be desirable to generate two or more procollagens from distinct pro-α chains of an exogenous source in a host cell in which case it is required that co-assembly of pro-α chains to form undesirable hybrid molecules should not occur.

It is also conceivable that procollagens may need to be assembled in a cell-free system *in vitro*, in which case co-assembly of pro-α chains giving rise to undesirable hybrid molecules also needs to be avoided

It is an object of the present invention to provide a means by which pro-α chains or derivatives thereof may be assembled into desired procollagens or derivatives thereof without undesirable co-assembling with other pro-α chains.

According to the present invention there is provided a method of producing a desired procollagen or derivative thereof in a system which co-expresses and assembles at least one further procollagen or derivative thereof wherein the gene(s) for expressing pro-α chains or derivatives thereof for assembly into the desired procollagen has or have been exogenously selected from natural pro-α chains or exogenously manipulated such as to express said pro-α chains or derivatives thereof with domians which have the activity of C- terminal propeptide domains but which will not co-assemble with the C- terminal propeptide of the pro-α chains or derivatives thereof that assemble to form the said at least one further procollagen or derivative thereof.

By "procollagen or derivative thereof" and "pro-α chain or derivative thereof" we mean molecules of procollagen or pro-α chains respectively that may be identical to those found in nature or may be non-natural derivatives which may be proteins or

derivatives of proteins. Non-natural derivatives may also have non-protein domains or even be entirely a non-protein provided that the derivative contains a domain with activity of a C- terminal propeptide domain which will not co-assemble with the C-terminal propeptide domains of the pro-$\alpha$ chains or derivatives thereof that assemble to form at least one further procollagen or derivative thereof.

Preferred pro-$\alpha$ chain derivatives comprise a domain with the activity of a C-terminal propeptide domain and a further domain which is at least partially capable of trimerising to triple helix.

Thus the exogenously selected or exogenously manipulated genes may express pro-$\alpha$ chains or derivatives thereof that may be assembled into trimers to form procollagen molecules or derivatives thereof, which in turn may be formed into collagen polymers following exposure to Procollagen C-Proteinase and Procollagen N-Proteinases (which respectively cleave the C- and N- terminal propeptides from the procollagen molecules to form monomers which aggregate spontaneously to form the collagen polymers). The collagen polymer is preferably a fibrillar collagen

The invention is based upon the recognition by the inventors that a crucial stage in the assembly of procollagens is an initial recognition step between pro-$\alpha$ chains which ensures that pro-$\alpha$ chains assemble in a type-specific manner. This recognition step involves a recognition sequence in the C- terminal propetide domain of pro-$\alpha$ chains. For instance, a single cell may synthesise several collagen types and, therefore, several different pro-$\alpha$ chains, yet these chains are able to discriminate between C- terminal propetide domains to ensure type-specific assembly. One example of this discrimination can be found in cells expressing both type I and type III procollagen. Here at least three pro-$\alpha$ chains are synthesised, namely pro$\alpha$1(I), pro$\alpha$2(I) and pro$\alpha$1(III) chains. However the only procollagens formed are [pro$\alpha$1(I)]$_2$pro$\alpha$2(I) heterotrimers and [pro$\alpha$1(III)]$_3$ homotrimers. Other combinations of pro-$\alpha$ chains do not assemble into procollagens.

5

In PCT/GB96/02122 (WO-A-97/08311) the disclosure of which is incorporated by reference we have disclosed that specific regions within the C-terminal propeptide are the recognition sequences involved in the specificity of association between C-terminal propeptide domains of pro-$\alpha$ chains during the formation of procollagens. These recognition sequences were identified as having the following amino acid sequences for each respective pro-$\alpha$ chain:

| | |
|---|---|
| pro-$\alpha$1(I) | GGQGSDPADV AIQLTFLRLL M STE |
| pro-$\alpha$2 (I) | NVEGVTSKEM ATQLAFMRLL ANY |
| pro-$\alpha$1 (II) | GDDNLAPNTA NVQMTFLRLL STE |
| pro-$\alpha$1 (III) | GNPELPEDVL DVQLAFLRLL SSR |
| pro-$\alpha$1 (V) | VDAEGNPVGV VQMTFLRLL SAS |
| pro-$\alpha$2 (V) | GDHQSPNTAI TQMTFLRLL SKE |
| pro-$\alpha$1 (XI) | LDVEGNSINM VQMTFLKLL TAS |
| pro-$\alpha$2 (XI) | VDSEGSPVGV VQLTFLRLL SVS |

These recognition sequences confer selectivity and specificity of pro-$\alpha$ chain association.

In accordance with the invention, we have devised methods by which desired pro-$\alpha$ chains or derivatives thereof can be expressed and assembled into procollagens or derivatives thereof in a system which co-expresses and assembles pro-$\alpha$ chains or derivatives thereof of at least one further procollagen or derivative thereof without undesired co-assembly producing unwanted hybrid molecules. This is effected by exogenously manipulating or selecting the gene or genes that encode for the desired pro-$\alpha$ chains or derivatives thereof such that the domains having C-terminal propeptide activity of these pro-$\alpha$ chains or derivatives thereof that are expressed from the manipulated or selected gene or genes will not associate with (and therefore not co-assemble with) the domains having C-terminal propeptide activity of the pro-$\alpha$

6

chains or derivatives thereof of the said at least one further procollagen or derivative thereof. Put alternatively, the domains having C- terminal propeptide activity of the pro-α chain or derivative expressed by the manipulated or selected gene are such that association between pro-α chains expressed from such a gene and association between at least one pro-α chain which forms the further procollagen or derivative thereof is mutually exclusive.

Thus, in accordance with the present invention, a gene for expressing a pro-α chain or derivative thereof for assembly into a desired procollagen may be exogenously selected or constructed to express a pro-α chain or derivative thereof comprised of (i) a first moiety incorporating at least the recognition sequence of the C-terminal propeptide domain of a first type of pro-α chain, and (ii) a second moiety, attached to the first moiety which will assemble into the desired procollagen. The second moiety preferably is at least partially capable of trimerising to form a triple helix. More preferably the second moiety comprises at least some amino acids capable of trimerising with other α chains or derivatives thereof. The expressed molecule is one which has been "engineered" (by appropriate selection of the first and second moieties) such that it may be expressed and assembled in a system which co-expresses and assembles at least one further type of pro-α chain without undesirable formation of hybrid molecules.

The domain having C-propeptide activity expressed by the exogenously selected or modified gene may comprise a recognition sequence as listed above. The domain may be a modification (e.g. by substitution or deletion) of such a recognition sequence, the domain retaining C-propeptide activity.

To prepare exogenously modified genes for use in the method of the invention, the DNA encoding for the desired recognition sequence may be substituted for the DNA encoding recognition sequences found in natural or artificially constructed pro-

7

α chain genes to form an exogenously modified gene for use in the method of the invention.

DNA, particularly cDNA, encoding natural pro-α chains is known and available in the art. For example, WO-A-9307889, WO-A-9416570 and the references cited in both of them give details. Such DNA may be used as a convenient starting point for making a DNA molecule that encodes for an exogenously manipulated gene for use in the invention.

DNA sequences, cDNAs, full genomic sequences and minigenes (genomic sequences containing some, but not all, of the introns present in the full length gene) may be inserted by recombinant means into a DNA sequence coding for naturally occurring pro-α chains (such as the starting point DNA mentioned above) to form the DNA molecule that encodes for an exogenously manipulated gene for use according to the first aspect of the invention.. Because of the large number of introns present in collagen genes in general, experimental practicalities will usually favour the use of cDNAs or, in some circumstances, minigenes. The inserted DNA sequences, cDNAs, full genomic sequences or minigenes code for amino acids which give rise to pro-α chains or derivative thereof with a C- terminal propeptide domain which will not co-assemble with the C- terminal propeptide domain of the pro-α chains or derivatives thereof that assemble to form the said at least one further procollagen or derivative thereof.

Preferred exogenous manipulations of the gene or genes involve alteration of the recognition sequence within the C- terminal propeptide domain which is responsible for selective association of pro-α chains such that any pro-α chain or derivative thereof expressed from the manipulated gene will not undesirably co-assemble with pro-α chains endogenously expressed from a host cell into which the exogenously manipulated gene or genes is or are introduced.

8

In our previous application PCT/GB96/02122 (WO-A-97/08311) we disclosed novel molecules comprising combinations of natural or novel C- terminal propeptide domains with alien α chains (or a non-collagen material). PCT/GB96/02122 also disclosed DNA molecules encoding such molecules. These DNA molecules may be used according to the methods of the current invention. Such molecules disclosed in PCT/GB96/02122 are incorporated herein by reference.

Alternatively deletion, addition or substitution mutations may be made within the DNA encoding for any one of these recognition sequences which alter selectivity and specificity of pro-α chain association.

Other preferred exogenous manipulations of a gene involve the construction of gene constructs which encode for chimeric pro-α chains or derivatives thereof formed from the genetic code of at least two different pro-α chains. It is particularly preferred that the chimeric pro-α chains or derivatives thereof comprise a recognition sequence from the C- terminal propeptide domain of one type of pro-α chain and the α chain domain from another type of pro-α chain. Preferred chimeric pro-α chains or derivatives thereof comprise the recognition sequence of a pro-α1(I), pro-α2 (I), pro-α1 (II), pro-α1 (III), pro-α1 (V), pro-α2 (V), pro-α1 (XI) or pro-α2 (XI) pro-α chain and an α-chain domain selected from a different one of these pro-α chains. Most preferred pro-α chains for making chimeric pro-α chains or derivatives thereof are those which form collagens I and III particularly pro-α2 (I) and pro-α1 (III). Specific preferred chimeric pro-α chains or derivatives thereof are disclosed in the Example.

In a preferred exogenous manipulation of a gene according to the methods of the invention, the DNA encoding for the recognition sequence of the proα2(I) chain gene can be replaced with the corresponding DNA encoding for the recognition sequence of the proα1(III) chain gene and this manipulated gene can be expressed and assembled to form procollagens which are proα2(I) homotrimers (instead of

proα1(III) homotrimers which would normally be formed from pro-α chains containing these recognition sequences). Thus according to the invention proα2(I) homotrimers derived from an exogenous source may be formed which do not co-assemble with proα2(I) chains endogenous to the cell in which expression occurs which have "natural" recognition sequences.

In another preferred exogenous manipulation of a gene according to the methods of the invention, the manipulated gene encodes for a molecule comprising at least a first moiety having the activity of a procollagen C-propeptide (i.e. the C-terminal propeptide domain of a pro-α chain) and a second moiety selected from any one of an alien collagen α chain and non-collagen materials, the first moiety being attached to the second moiety. Genes which encode for a second moiety of a non-collagen material (such as those disclosed in PCT/GB96/02122) are examples of pro-α chain derivatives for use according to the invention.

Alternatively the gene or genes may be selected from naturally occurring genes such that the recognition sequence within the C- terminal propeptide domain which is responsible for selective association of pro-α chains such that any pro-α chain expressed from the selected gene will not undesirably co-assemble with pro-α chains endogenously expressed from the host cell into which the gene or genes is or are introduced.

The exogenously selected or modified gene may be incorporated within a suitable vector to form a recombinant vector. The vector may for example be a plasmid, cosmid or phage. Such vectors will frequently include one or more selectable markers to enable selection of cells transfected with the said vector and, preferably, to enable selection of cells harbouring the recombinant vectors that incorporate the exogenously modified gene.

For expression of pro-α chains or derivatives thereof the vectors should be expression vectors and have regulatory sequences to drive expression of the exogenously modified gene. Vectors not including such regulatory sequences may also be used during the preparation of the exogenously modified gene and are useful as cloning vectors for the purposes of replicating the exogenously modified gene. When such vectors are used the exogenously modified gene will ultimately be required to be transferred to a suitable expression vector which may be used for production of the pro-α chains or derivatives thereof.

The system in which the exogenously selected pro-α chain(s) or exogenously manipulated gene or genes of the method of the invention may be expressed and assembled into procollagen or derivatives thereof may be a cell free *in vitro* system. However it is preferred that the system is a host cell which has been transfected with a DNA molecule according to the second aspect of the invention. Such host cells may be prokaryotic or eukaryotic. Eukaryotic hosts may include yeasts, insect and mammalian cells. Hosts used for expression of the protein encoded by the DNA molecule are ideally stably transformed, although the use of unstably transformed (transient) hosts is not precluded.

Alternatively a host cell system may involve the DNA molecule being incorporated into a transgene construct which is expressed in a transgenic plant or, preferably, animal. Transgenic animals which may be suitably formed for expression of such transgene constructs, include birds such as domestic fowl, amphibian species and fish species. Procollagens or derivatives thereof and / or collagen polymers formed therefrom may be harvested from body fluids or other body products (such as eggs, where appropriate). Preferred transgenic animals are (non-human) mammals, particularly placental mammals. An expression product of the DNA molecule of the invention may be expressed in the mammary gland of such mammals and the expression product may subsequently be recovered from the milk. Ungulates, particularly economically important ungulates such as cattle, sheep, goats, water

buffalo, camels and pigs are most suitable placental mammals for use as transgenic animals according to the invention. Equally the transgenic animal could be a human in which case the expression of the pro-α chains or derivative thereof in such a person could be a suitable means of effecting gene therapy.

Host cells and particularly transgenic plants or animals, may contain other exogenous DNA, the expression of which facilitates the expression, assembly, secretion or other aspects of the biosynthesis of procollagen and derivatives thereof and even collagen polymers formed therefrom. For example, host cells and transgenic plants or animals may also be manipulated to co-express prolyl 4-hydroxylase, which is a post translation enzyme important in the natural biosynthesis of procollagens, as disclosed in WO-A-9307889.

The methods of the invention enable the expression and assembly of any desired procollagen or derivative thereof in a system in which conventionally there would be undesirable co-assembly or hybridisation of pro-α chains. The methods are particularly suitable for allowing the expression of procollagen or derivatives thereof from a wide variety of cell-lines or transgenic organisms without the problems associated with co-assembly with endogenously expressed pro-α chains. A preferred use of the methods of the invention is the production of recombinant procollagens in cell-lines. Examples of cell-lines which may be used are fibroblasts or cell lines derived therefrom. Baby Hamster Kidney cells (BHK cells), Mouse 3T3 cells, Chinese Hamster Ovary cells (CHO cells) and COS cells may be used.

The methods of the invention are particularly useful as an improved means of production of any desired procollagen or derivatives thereof, particularly for scaled up industrial production by biotechnological means.

The method of the invention may also be useful for treatment by gene therapy of patients suffering from diseases such as osteogenesis imperfecta (OI), some forms

12

of Ehlers-Danlos syndrome (EDS) or certain forms of chrondrodysplasia. In most cases the devastating effects of these diseases are due to substitutions of glycine within the triple helical domain, for amino acids with bulkier side chains in the pro-α chains. This substitution results in triple helix folding, during the formation of procollagen, being prevented or delayed with the consequence that there is a drastic reduction in the secretion of the procollagen. The malfolded proteins are retained within the cell, probably within the endoplasmic reticulum, where they are degraded. Furthermore, the folding of the C-terminal propeptide domain is not affected by these mutations within the triple helical domain, therefore C-terminal propeptide domains from normal as well as mutant chains may associate resulting in the retention of normal and mutant pro-α chains within the cell. The retention and degradation of normal chains due to their interaction with mutant chains amplifies the effect of the mutation and has been termed "procollagen suicide". The massive loss of protein due to this phenomenon probably explains why such mutations produce lethal effects. Identification by the inventors of the recognition sequence which directs the initial association between pro-α chains provides a target for therapeutic intervention allowing for the modulation or inhibition of collagen deposition. Thus, the method of the invention could be utilised as a gene therapy to transfer a copy of the wild-type gene to an individual with a mutation in the triple helical domain such that the wild-type gene is exogenously manipulated to code for a pro-α chain with a C-terminal propeptide domain that will not co-assemble with the mutant pro-α chains. The patient is then able to secrete authentic collagen chains in cells expressing mutant chains

The present invention will now be described, by way of example with reference to the accompanying drawings, in which:

Figure 1 is a schematic representation of the stages in normal procollagen assembly (A) and stages in procollagen assembly according to one embodiment of the invention (B):

Figure 2 shows an alignment plot of the C- terminal propeptide domains of pro-α chains from type I and III collagen. The alignment shows amino acids which are identical (#) or those which are conserved (-). The conserved cysteine residues are numbered 1-8, while letters A, B, C, F, G denote the first amino acid at the junctions between proα1(III) chains and proα2(I) chains of the Example:

Figure 3 is a schematic representation of the chimeric pro-α1 chains described in the Example:

Figure 4 is a photograph of an SDS-PAGE gel, illustrating disulphide bond formation among chimeric gene constructs in which the C-terminal propeptide domain were exchanged, with the following parental and chimeric molecules from the Example run in the indicated lanes of the gel. Proα1 (III)A1 [α1(III)], proα2(I)A1 [α2(I)] (parental molecule) and proα2(I):(III)CP [α2 CP], proα1(III):(I)CP [α1:CP] (hybrid chains). these molecules were expressed in a rabbit reticulocyte lysate in the presence of semi-permeabilized (SP) HT 1080 cells, after which the SP-cells were isolated by centrifugation, solubilized and the translation products separated by SDS-PAGE through a 7.5% gel under reducing (lanes 1-4) or non-reducing conditions (lanes 5-8):

Figure 5 is a photograph of an SDS-PAGE gel the lanes represent the effect of heat denaturation of proα2(I):(III)CP triple-helix at the specified temperatures. the samples were prepared in the following manner: Proα2(I):(III)CP RNA was translated in the presence of SP-cells, after which the SP-cells were isolated by centrifugation, solubilized and treated with pepsin (100 µg/ml). the reaction mixture was neutralized, diluted in chymotrypsin/trypsin digest buffer and divided into aliquots. each aliquot being heated to a set temperature prior to digestion with a combination of trypsin (100 µg/ml) and chymotrypsin (250 µg/ml). samples were analysed by SDS-PAGE through

14

a 12.5% gel under reducing conditions (lanes 1-10). Lane 11 (unt) contains translation products which have not been treated with proteases;

Figure 6 is a photograph of an SDS-PAGE gel illustrating trimerization and triple-helix formation among chimeric procollagen chains, samples were prepared from parental chains $pro\alpha1(III)\Delta1$, $pro\alpha2(I)\Delta1$ which were made into hybrids $pro\alpha2(I):(III)CP$, $A.F.F^{S-C}$, $Pro\alpha1(III):(I)C$ ($\alpha2CP$, $A.F.F^{S-C},B^{S-C},C^{S-C}$, $\alpha1C$). the hybrids were translated in a rabbit reticulocyte lysate in the presence of SP-cells after which the SP-cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS-PAGE under non-reducing conditions through a 7.5% gel (lanes 1-9).

Figure 7 is a photograph of an SDS-PAGE gel illustrating trimerization and triple-helix formation among chimeric procollagen chains. lanes show the remainder of the samples that were loaded on the gel of Fig 6 which were treated with pepsin (100 µg/ml) prior to neutralization and digestion with a combination of trypsin (100 µg/ml) and chymotrypsin (250 µg/ml), the proteolytic digestion products were analysed by SDS-PAGE through a 12.5% gel under reducing conditions (lanes 1-9);

Figure 8 is a photograph of an SDS-PAGE gel, illustrating trimerization and triple-helix formation among chains containing the 23 amino acid B-G motif, the lanes show recombinant procollagen chains $pro\alpha1(III):(I)CP$, $pro\alpha2(I):(III)CP$ and $pro\alpha2(I):(III)BGR^{S-C}$ which were expressed in a reticulocyte lysate supplemented with SP-cells, after which the SP-cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS-PAGE through a 7.5% gel, under reducing (lanes 1-3) of non-reducing conditions (lanes 4-5).

Figure 9 is a photograph of an SDS-PAGE gel, illustrating trimerization and triple-helix formation among chains containing the 23 amino acid B-G motif, the lanes show the remainder of the samples that were loaded on the gel of Figure 9 which

were treated with pepsin (100 μg/ml) prior to neutralization and digestion with a combination of trypsin (100 μg/ml) and chymotrypsin (200 μg/ml), the proteolytic digestion products were analysed by SDS-PAGE through a 12.5% gel under reducing conditions (lanes 1-3);

Figure 10 is a photograph of an SDS-PAGE gel, illustrating the effect of Cys-Ser reversion and Leu-Met mutation on the assembly of proα2(I):(III)BGR chains, the lane show recombinant procollagen chains proα2(I):(III)BGR$^{S-c}$ proα2(I):(III)BGR$^{c-s}$, proα2(I):(III)BGR$^{l-m}$ which were translated in a reticulocyte lysate supplemented with SP-cells after which the cells were isolated by centrifugation, solubilized and a portion of the translated material separated by SDS-PAGE through a 7.5% gel, under reducing (lanes 1-3) or non-reducing conditions (lanes 4-6);

Figure 11 is a photograph of an SDS-PAGE gel, illustrating the effect of Cys-Ser reversion and Leu-Met mutation on the assembly of proα2(I):(III)BGR chains, the lane show the remainder of the samples that were loaded on the gel of Fig 10 which were treated with pepsin (100 μg/ml) prior to neutralization and digestion with a combination of trypsin (100 μg/ml) and a chymotrypsin (250 μg/ml), the proteolytic digestion products were analysed by SDS-PAGE through a 12.5% gel under reducing conditions (lanes 1-3);

Figure 12 is a photograph of an SDS-PAGE gel, illustrating inter-chain disulfide bonds from between proα2(I):(III)BGR C-terminal propeptide domains, the lanes show recombinant pro-α chains proα1(III)A1 and proα2(I):(III)BGR which were translated in a reticulocyte lysate supplemented with SP-cells. The cells were isolated by centrifugation, solubilized and digested with 1.5 units of bacterial collagenase. The products of digestion were analysed by SDS-PAGE through a 10% gel under reducing (lanes 2 and 3) or non-reducing (lanes 4 and 5) conditions; and

Figure 13 is a schematic representation of sequence alignment of the chain selectivity recognition domains in other fibrillar procollagens, sequence homology within the 23 residue B-G motif is illustrated, the boxed regions indicating the position of the unique 15 residue sub-domain which directs pro-α chain discrimination.

Figure 1 illustrates how procollagen is assembled in the endoplasmic reticulum of a cell. Normally assembly is initiated by type specific association of C-terminal propeptide domains of complimentary pro-α chains (1) to form procollagens (2). Procollagen is secreted from the cell in which it is synthesised and is then acted upon by Procollagen N Proteinases and Procollagen C Proteinases which cleave the N-terminal propeptide and C-terminal propeptide respectively to yield collagen molecules (3). Collagen molecules may then spontaneously aggregate to form collagen fibrils. Pro-α chains with non-complimentary C-terminal propeptide domains (4) do not associate and form procollagens. When exogenous pro-α chains (5) are introduced into a cell they may co-assemble with endogenous pro-α chains (6) which have complimentary C-terminal propeptide domains to form undesirable hybrids (7). According to the methods of the invention exogenously manipulated pro-α chains (8) are generated with C- terminal propeptide domains that are no longer complimentary to the C-terminal propeptide domains of the endogenous pro-α chains (6) such that the exogenously manipulated pro-α chains (8) may form procollagens (9) and subsequently collagen molecules (10) without co-assembly with endogenous pro-α chains (6) occurring.

EXAMPLE

The inventors generated DNA molecules which may be used according to the methods of the invention. These DNA molecules were used to express pro-α chains with altered selectivity for pro-α chain assembly. Experimental strategy was based on the assumption that transfer of C- terminal propeptide domains (or sequences within the C-propeptide) from the homotrimeric proα1(III) chain to the proα2(I) molecule would be sufficient to direct self-association and assembly into homotrimers of proα2(I). The inventors reconstituted the initial stages in the assembly of procollagen by expressing specific RNAs in a cell-free translation system in the presence of semi-permeabilized cells known to carry out the co- and post-translational modification required to ensure assembly of a correctly aligned triple helix. By analysing the folding and assembly pattern of procollagens formed from a series of chimeric pro-α chains in which specific regions of the C-terminal propeptide domain of proα1 (III) were exchanged with the corresponding region within the proα2(I) chain (and vice versa) the inventors identified a short discontinuous sequence of 15 amino acids within the proα1 (III) C-propeptide which directs procollagen self-association. This sequence is, therefore, responsible for the initial recognition event and is necessary to ensure selective chain association.

1.      MATERIALS AND METHODS

### 1.1 Construction of recombinant plasmids

pα1(III)Δ1 and pα2(I)Δ1 are recombinant pro-α chains with truncated α chain domains which have been described previously (see Lees and Bulleid (1994) J. Biol Chem. 269 p24354-24360 1994). Chimaeric molecules were generated by PCR overlap extension using the principles outlined by Horton (1993) Methods in Molecular Biology Vol 15, Chapter 25, Humana Press Inc., Totowa, NJ. PCRs (100μl) compromised template DNA (500 ng), oligonucleotide primers (100 pmol each) in 10 mM KCl, 20 mM Tris-HCl pH 8.8, 10mM $(NH_4)_2SO_4$, 2 mM $MgSO_4$, 0.1% (v/v) Triton X-100, 300 μM each dNTP. Ten rounds of amplification were

18

performed in the presence of 1 unit Vent DNA polymerase (New England Biolabs, MA). Recombinants pα2(I)Δ1:(III)CP, A, F, S$^{xx}$, C$^{xxx}$ were generated using a 5' oligonucleotide primer (5 'AGATGGTCGCACTGGACATC 3') complementary to a sequence 70 bp upstream of an *Sfil* site in pα2(I)Δ1 and a 3' oligonucleotide primer (5' TCGCAGGGATCCGTCGGTCACTTGCACTGGTT 3') complementary to a region 100 bp downstream to the stop codon in pα1(III)Δ1. A *Bam*HI site was introduced into this primer to facilitate subsequent sub-cloning steps. Pairs of internal oligonucleotides, of which one included a 20 nucleotide overlap, were designed to generate molecules with precise junctions as delineated (see Figs 2 and 3) Overlap extension yielded a product of ~990 bp which was purified, digested with *Xhol-BamHI* and ligated into pα2(I)Δ1 from which a 1080 bp *Xhol-BamHI* fragment had been excised. Recombinants pα1(III)Δ1:(I)CP,C were synthesized in a similar manner using a 5' oligonucleotide (5' AATGGAGCTCCTGGACCCATG 3') complementary to a sequence 100bp upstream of an *Xhol* site in a pα(III)Δ1 and a 3' amplification primer (5 'CTGCTAGGTACCAAATGGAAGGATTCAGCTTT 3') which incorporated a *Kpnl* site and was complementary to a region 100bp downstream of the stop codon in pα2(I)Δ1. Overlap extension produced a fragment of 1100 bp which was digested with *Xhol* and *Kphl* and ligated into pα1(III)Δ from which an 1860 bp fragment had been removed. Recombinant pα2(I):(III)BGR was constructed using the same amplification primer used to synthesize the proα2(I)Δ1:(III) series of chimeras and a 3' oligonucleotide which was identical to that used to generate the proα1(III)Δ1:(I)CP,C constructs except that it contained a *Bam*HI site instead of *Kpnl* (both complementary to pα2(I)Δ1). Primary amplification products were generated from pα2(I)Δ1:(III)B$^{xx}$ and pα2(I)Δ1 with internal oligonucleotides determining the junction. Overlap extension produced a fragment which was digested with *Sfil* and *Bam*HI and ligated into pα2(I)Δ1. Site-directed mutagenesis was performed essentially as described by Kunkel *et al.* (Kunkel *et al.* (1987) Methods in Enzymol. 154 p 367-382), except that extension reactions were performed in the presence of 1 unit T4 DNA polymerase and 1 μg T4 gene 32 protein (Boehringer, Lewes, UK).

### 1.2 Transcription in vitro

Transcription reactions were carried out as described by Gurevich *et al.* (1987) (see Gurevich *et al.* (1991) Anal. Biochem. 195 p207-213) . Recombinant plasmids pα1(III)Δ1, pα1(III)Δ1:(I)CP,C and pα2(I)Δ1, pα2(I)Δ1:(III)CP, A, F, F\*\*, B\*\*, C\*\* (10µg) were linearized and transcribed using T3 RNA polymerase, or T7 RNA polymerase (Promega, Southampton, UK) respectively. Reactions (100 µl) were incubated at 37°C for 4 h. Following purification over RNeasy columns (Qiagen, Dorking, UK), RNA was resuspended in 100 µl RNasefree water containing 1 mM DTT and 40 units RNasin (Promega, Southampton, UK).

### 1.3 Translation in vitro

RNA was translated using a rabbit reticulocyte lysate (FlexiLysate, Promega, Southampton) for 2 hours at 30°C in the absence of exogenous DTT. The translation reaction (25µl) contained 17µl reticulocyte lysate, 1 µl 1 mM amino acids (minus methionine), 0.45 µl 100mM KCl, 0.25 µl ascorbic acid (5 mg/ml), 15 µCi [l-$^{35}$S]methionine (Amersham International, Bucks, UK), 1 µl transcribed RNA and 1 µl (~2 x 10$^5$) semi-permeabilized cells (SP-cells) prepared as described by Wilson *et al.* (1995) Biochem. J. 307 p679-687. After translation, *N*-ethylmaleimide was added to a final concentration of 20 mM. SP-cells were isolated by centrifugation in a microfuge at 10000 g for 5min and the pellet resuspended in an appropriate buffer for subsequent enzymic digestion or gel electrophoresis.

### 1.4 Bacterial collagenase digestion

SP-cells were resuspended in 50 mM Tris-HCl pH 7.4 containing 5 mM CaCl$_2$, 1 mM phenylmethanesulfonyl fluoride (PMSF), 5mM *N*-ethylmaleimide and 1% (v/v) Triton X-100 and incubated with 3 units collagenase form III (Advance Biofacture, Lynbrook, NJ) and incubated at 37°C for 1h. The reaction was terminated by the addition of SDS-PAGE sample buffer.

## 1.5 Proteolytic digestion

Isolated SP-cells were resuspended in 0.5% (v/v) acetic acid, 1% (v/v) Triton X-100 and incubated with pepsin (100 μg/ml) for 2 h at 20°C or 16 h at 4°C. The reactions were stopped by neutralization with Tris-base (100 mM). Samples were then digested with a combination of chymotrypsin (250 μg/ml) and trypsin (100 μg/ml) (Sigma, Poole, Dorset, UK) for 2 min at room temperature in the presence of 50 mM Tris-HCl pH 7.4 containing 0.15 M NaCl, 10 mM EDTA. The reactions were stopped by the addition of soy bean trypsin inhibitor (Sigma, Poole, Dorset, UK) to a final concentration of 500 μg/ml and boiling SDS-PAGE loading buffer. Samples were then boiled for 5 min.

## 1.6 Thermal denaturation

Pepsin-treated samples were resuspended in 50 mM Tris-HCl pH 7.4 containing 0.15 M NaCl, 10mM EDTA, and aliquots placed in a thermal cycler. A stepwise temperature gradient was set up from 31°C to 40°C with the temperature being held for 2 min at 1°C intervals. At the end of each time period the sample was treated with a combination of chymotrypsin, as described above.

## 1.7 SDS-PAGE

Samples resuspended in SDS-PAGE loading buffer (0.0625 M Tris-HCl pH 6.8, SDS (2% w/v), glycerol (10% v/v) and Bromophenol Blue) in the presence or absence of 50 mM DTT and boiled for 5 min. SDS-PAGE was performed using the method of Laemmli (1970) Nature 227 p680-685. After electrophoresis, gels were processed for autoradiography and exposed to Kodak X-Omat AR film, or images quantified by phosphoimage analysis.

## 2. RESULTS

### 2.1 Transfer of the proα1(III) C-propeptide to the proα(I)2 chain is sufficient to direct self-assembly.

Experimental strategy was based on the assumption that transfer of the C-terminal propeptide domain from the proα1(III) chain to the proα2(I) chain should be sufficient to direct self-recognition and assembly into homotrimers. Hence, by exchanging different regions within the proα1(III) C-terminal propeptide domain with the corresponding sequence from the proα2(I) chain the intention was to distinguish between sequences that direct the folding of tertiary structure and those involved in the selection (i.e. recognition of pro-α chains) process. To simplify analysis of the translation products chimeric procollagen molecules were constructed from two parental procollagen 'mini-chains', proα1(III)Δ1 and proα(I)Δ1. These molecules, which have been described previously (Lees and Bulleid, 1994), comprise both the N- and C-terminal propeptides domains together with truncated triple-helical domains. The initial assumption was tested by analysing the folding and assembly of chimeric procollagen chains in which the C-terminal propeptide domain of the proα2(I) chain was substituted with the equivalent domain from the proα1(III)Δ1 chain (proα2-(I):(III)CP) and, conversely, where the C-propeptide of proα1(III) chain was replaced with that from proα2(I)Δ1 chain (proα1(III):(I)CP) (see Figs 2 and 3). The C-propeptide (CP) junction points were determined by the sites of cleavage by the procollagen C-proteinase (PCP) which is known to occur between Ala and Asp (residues 1119-1120) in the proα2(I) chain (Kessler (1996) Science 271 p360-362). In the absence of data regarding the precise location of cleavage within the proα(III) chain, the inventors chose to position the junction between Ala and Pro (residues 1217-1218). However, Kessler and co-workers (1996) have subsequently shown that cleavage by PCP occurs between Gly and Asp (residues 1222-1223), with the consequence that recombinant proα2(I):(III)CP includes an additional four residues derived from the proα(III) C-telopeptide, whilst the C-telopeptide in construct proα1(III):(I)CP is missing those same four amino acids. RNA transcripts were

transcribed *in vitro* and expressed in a cell-free system comprising a rabbit reticulocyte lysate optimized for the formation of disulfide bonds supplemented with semi-permeabilized HT 1080 cells (SP-cells), which has been shown previously to carry out the initial stages in the folding, post-translational modification and assembly of procollagen (Bulleid *et al.*, (1996) Biochem. J. 317 p195-202). The C-terminal propeptide domains of both proα1(III) and proα2(I) chains contain cysteine residues which participate in the formation of interchain disulfide bonds. Translation products were, therefore, separated by SDS-PAGE under reduced and non-reduced conditions in order to detect disulfide-bonded trimers. Translation of the parental molecules proα1(III)Δ1 and proα2(I)Δ1 yielded major products of ~77 kDa and 61 kDa respectively (Figure 4, lanes 1 and 2), the size differential being accounted for by the relative molecular weights of the N-propeptides and truncated triple-helical domains in each molecule (Lees and Bulleid, 1994). The heterogeneity of the translation products is due to hydroxylation of proline residues in the triple-helical domain that leads to an alteration in electrophoretic mobility (Cheah *et al.*, (1979) Biochem. Biophys. Res. Comm. 91 p1025-1031). The additional lower molecular weight proteins present in lanes 3 and 7 probably represent translation products obtained after initiation of translation at internal start codons. We have previously shown that these minor translation products are not translocated into the endoplasmic reticulum (Lees and Bulleid, 1994). The presence of high molecular weight species under non-reducing conditions but not reducing conditions is indicative of interchain disulfide bond formation. Separation under non-reduced conditions revealed that proα1(III)Δ1, but not proα2(I)Δ1, chains were able to self-associate to form disulfide-bonded trimers (Figure 4, lanes 5 and 6). A similar examination of chimeric chains proα2(I):(III)CP and proα1(III):(I)CP revealed that only proα2(I):(III)CP chains were able to form disulfide-bonded homotrimers (Figure 4, lanes 3, 4, 7 and 8) demonstrating that the C-propeptide from type III procollagen is both necessary and sufficient to drive the initial association between procollagen chains.

It has been shown previously that proα1(III)Δ1 chains synthesised in the presence of SP-cells were resistant to a combination of pepsin, chymotrypsin and trypsin in a standard assay used specifically to detect triple-helical procollagen (Bulleid et al., 1996). The inventors confirmed that proα2(I):(III)CP chains had the ability to form a correctly aligned triple-helix by performing a thermal denaturation experiment in which translated material was heated to various temperatures prior to protease treatment (Figure 5). The results indicate that at temperatures below 35°C a protease-resistant triple-helical fragment is present, but at temperatures above 35°C the triple-helix melts and becomes protease sensitive (Figure 5, lanes 1-10). The melting temperature ($T_m$) was calculated to be ~35.5°C after quantification by phophorimage analysis. The $T_m$ value obtained for proα2-(I):(III)CP is significantly lower than the figure of 39.5°C obtained for proα1(III)Δ1 (Bulleid et al., 1996) and probably reflects the percentage of hydroxyproline residues relative to the total number of amino acids in the triple-helical domain (11% and 15% respectively). These results indicate that transfer of the proα(III) C-propeptide enables the inventors to generate an entirely novel procollagen species comprising three proα2(I) chains that fold into a correctly aligned triple-helix.

### 2.2 Assembly of recombinant procollagen chains with chimeric C-propeptides.

Given that the proα2(I):(III)CP hybrid pro-α chain includes all of the information required for self-association we reasoned that progressive removal of the proα1(III) C-propeptide sequence and replacement with the corresponding proα2(I) sequence would eventually disrupt the chain selection mechanism. Conversely, it is anticipated that transfer or progressively more proα1(III) C-terminal propeptide domain sequence to the proα1(III):(I)CP chimeric chain would yield a molecule which was capable of self-assembly. A series of procollagen chains with chimeric C-terminal propeptide domains was constructed and the ability of individual chains to form homotrimers with stable triple-helical domains was assessed. A schematic representation of these recombinants is presented in Figure 2, with the letters A, B, C, F and G denoting the position of each junction. It should be noted that the proα1(III)

24

and proα2(I) C-propeptides differ in their complement of cysteine residues, with proα2(I) lacking the Cys2 residue. Our previous data suggest that interchain disulfide bond within the C-propeptide of type III procollagen form exclusively between Cys2 and 3 (Lees and Bulleid, 1994). However, interchain disulfide bonding, between either the C-terminal propeptide domains to C-telopeptides is not required for chain association and triple-helix formation (Bulleid et al., 1996), therefore, it is possible that homotrimers may form between chimeric pro-α chains which lack either the C-terminal propeptide domain Cys2 residue or the C-telopeptide cysteine [only found in the triple-helical domain of proα1(III)]. These molecules will not, however, contain interchain disulfide bonds and, as a consequence will not appear as oligomers after analysis under non-reducing conditions. To circumvent this problem, where appropriate, the inventors generated their hybrid chains from a recombinant proα2(I)A1"" (Lees and Bulleid, 1994) in which the existing serine residue was substituted for cysteine, thus restoring the potential to form trimers stabilized by interchain disulfide bonds. It should also be noted that whilst proα1(III):(I)CP lacks Cys2, it does still retain the potential to form disulfide-bonded trimers by virtue of the two cysteine residues located at the junction of the triple-helical domain and the C-telopeptide. Parental chains proα2(I)A1 and hybrids proα2(I):(III)CP, A, 1", F"", B"", C"", proα1(III):(I)C were translated in the presence of SP-cells and the products separated by SDS-PAGE under non-reducing conditions (Figure 6). The results demonstrate that recombinants proα1(III)Δ1, proα2(I):(III)CP, A, F"", B"" (Figure 6 lanes 1, 3, 4, 6 and 7) are able to form interchain disulfide-bonded trimers and dimers while proα1(III)Δ1, proα2(I):(III)1, C"" and proα1(III):(I)C (Figure 6, lanes 2, 5, 8 and 9) remain monomeric. We have already demonstrated that interchain disulfide bonding is not a prerequisite for triple-helix formation (Bulleid et al., 1996), therefore, the inability to form disulfide-bonded trimers does not preclude the possibility that the molecules assemble to form a triple-helix. To ascertain whether the chimeric chains had the ability to fold into a correctly aligned triple-helix, we treated translation products with a combination of pepsin, chymotrypsin and trypsin and analysed the digested material under reducing conditions by SDS-PAGE. As shown in Figure 7,

25

recombinants proα1(III)Δ1, proα2(I):(III)CP, A, F$^{s,c}$, F, B$^{s,c}$ (Figure 7, lanes 1, 3, 4, 5, 6 and 7) all yielded protease-resistant fragments. The size differential reflects the relative lengths of the triple-helical domains in each of the parental molecules [proα2(I)Δ1-185 residues and proα1(III)Δ1-192 residues]. The ability of proα2(I):(III)F to form a stable triple-helix confirms that interchain disulfide bonding is not necessary for triple-helix folding. Thus, hybrid molecules containing sequences from the proα2 C-terminal propeptide domains between the propeptide cleavage site and the B-junction are able to form homotrimers with stable triple-helical domains and, therefore, contain all of the information necessary to direct chain self-assembly. These results indicate that the signal(s) which controls chain selectivity must be located between the B-junction and the C-terminus of the C-propeptide. Neither proα2(I):(III)C$^{s,c}$ nor proα1(III):(I)C chains are able to fold into a triple helix. The inability of these reciprocal constructs to self-associate suggests that chain selectivity is mediated, either by a co-linear sequence that spans the C-junction or by discontinuous sequence domains located on either side of the C-junction.

### 2.3 Identification of a sequence motif from the proα1(III) C-propeptide which directs chain self-assembly

Procollagen chain selectivity is probably mediated through one or more of the variable domains located within the C-terminal propeptide domain. The sequence between the B- and C-junctions is one of the least conserved among the procollagen C-propeptides (Figure 2), yet to inventors have demonstrated that inclusion of this domain, in the absence of proα1(III) sequence distal to the C-junction, is not sufficient to direct chain assembly. To ascertain whether the recognition sequence for chain recognition had indeed been interrupted a further recombinant, proα2(I):(III)BGR$^{s,c}$ (B-G replacement) was generated, which contained all of the proα(I)Δ1 sequence apart from the Ser→Cys mutation at Cys2 and a stretch of 23 amino acids derived from the type III C-propeptide which spans the C-junction from points B to G, the B-G motif: $^b$GNPELPEDVLDV$^c$QLAFLRLLSSR$^g$ (underscoring indicates the most divergent residues, see Figure 2). The location of the G-boundary

in the replacement motif allowed for the inclusion of the first non-conserved residues after the C-junction (SR). When expressed in the presence of SP-cells the chimeric proα2(I):(III)BGR$^{**}$ chains were able to form inter-chain disulfide-bonded molecules (Figure 8, lane 6) demonstrating that the C-terminal propeptide domains were capable of self-association. Furthermore, this hybrid was able to fold and form a stable triple-helix as judged by the formation of a protease-resistant fragment (Figure 9, lane 3). Proα2(I):I):(III)BGR$^{**}$ contains a Ser→Cys substitution which enabled the inventors to assay for the formation of disulfide-bonded trimers. Previous data demonstrated that this substitution alone does not enable wild-type proα2(I)Δ1 claims to form homotrimers (Lees and Bulleid, 1994). Nevertheless, to eliminate the possibility that this mutation influences the assembly pattern a revertant proα(I):(III)BGR$^{**}$ which contains the wild-type complement of Cys residues was created. As expected proα2(I):(III)BGR$^{**}$ was unable to form disulfide-bonded trimers (Figure 10, lane 5) but did assemble correctly into a protease-resistant triple helix (Figure 11, lane 3). Thus, the 23-residue B-G motif contains all of the information required to direct procollagen self-assembly.

The ability of the proα2(I):(III)BGR$^{**}$ chains to form interchain disulfide bonds suggests that this molecules is able to associate via its C-propeptide. However, to confirm that this is indeed the case the inventors carried out a collagenase digestion of the products of the translation (Figure 12). Bacterial collagenase specifically digests the triple-helical domain, leaving both the N- and C- propeptides intact. The N-propeptides of both chains do not contain any methionine residues and as a consequence, the only radio labelled product remaining after digestion is the C-propeptide. Comparison of the samples separated under reducing and non-reducing conditions demonstrated that inter-chain disulfide-bonded trimers were formed within the C-terminal propeptide domains of proα1(III)Δ1 and proα2(I):(III)BGR$^{**}$ chains (Figure 12, lanes 2 and 4, and 3 and 5). This demonstrates that these chains do indeed associate via their C-terminal propeptide domains.

27

### 2.4 The effect of Leu→Met substitution on proα2(I):BGR assembly

Analysis of the 23 amino acid B-G motif from the proα1(III) and proα2(I) chains (Figure 13) indicates that residues 13-20 (QLAFLRLL) are identical with the exception of position 17, Leu (L) in proα1(III) and Met (M) in proα2(I). Using site-directed mutagenesis the inventors substituted the existing Leu residue with Met to create proα2(I):(III)BGR$^{l-m}$ and monitored the effect of this mutation on chain assembly. The Leu→Met mutagenesis was performed using recombinant proα(I):(III)BGR$^{wt}$ and proα2(I):(III)BGR$^{l-m}$ and were able to form interchain disulfide-bonded molecules when analysed under non-reducing conditions (Figure 10, lanes 4 and 6). Both constructs formed protease-resistant triple-helical domains (Figure 11, lanes 1 and 3). The Leu→Met substitution did not, therefore, disrupt the process of chain selection nor did it prevent the formation of a correctly aligned triple-helix. These observations lead to the conclusion that a discontinuous sequence of 15 amino acids: (GNPELPEDVLDV....SSR) contains all of the information necessary to allow procollagen chains to discriminate between each other and assemble in a type-specific manner.

### 3. DISCUSSION

The molecular mechanism which enables closely related procollagen chains to discriminate between each other is a central feature of the assembly pathway. The initial interaction between the C-terminal propeptide domains both ensures that the constituent chains are correctly aligned prior to nucleation of the triple-helix and propagation in a C- to N- direction, and that component chains associate in a collagen type-specific manner. As a consequence, recognition signals which determine chain selectivity are assumed to reside within the primary sequence of this domain, presumably within a region(s) of genetic diversity. By generating chimeric procollagen molecules from parental 'mini-chains' proα1(III)Δ1 and proα2(I)Δ1 the inventors have demonstrated that transfer of the proα1(III) C-terminal propeptide domain to the naturally hetrotrimeric proα2(I) molecule was sufficient to direct formation of homotrimers. Furthermore, analysis of a series of molecules in which

28

specific sequences were interchanged from proα1(III) and proα2(I) C-terminal propeptide domains allowed the inventors to identify a discontinuous sequence of 15 amino acids (GNPELPEDVLDV.......SSR) within the proα1(III) C-propeptide, which, if transferred to the corresponding region within the proα1(III) recognition motif to the proα2(I) chain did not appear to have an adverse effect on chain alignment, allowing the triple-helical domains to fold into a protease-resistant confirmation. This sequence motif is, therefore, both necessary and sufficient to ensure that procollagen chains discriminate between each other and assemble in a type-specific manner.

In order to establish a structure-function relationship for the chain recognition domain, the inventors examined the hydropathy profile and secondary structure potential of the 23-residue B-G sequence : GNPELPEDVLDVQLAFLRLLSSR. The data indicate that the 15-residue chain recognition motif: GNPELPEDVLDV....SSR is markedly hydrophilic, in contrast to the hydrophobic properties of the conserved region: QLAFLRLLL. These features are entirely consistent with a potential role for this motif in mediating the initial association between the component procollagen monomers. An examination of the 15-residue recognition motif from other fibrillar procollagens predicts that they are all relatively hyrophilic and probably assume a similar structural conformation, regardless of the degree of diversity in the primary sequence (Figure 13). It is, presumably, the nature of the amino acids changes which provides the distinguishing topographical features necessary to ensure differential chain association.    An examination of the B-G sequence alignment (Figure 13) indicates that residues 1, 2, 12 and 21 are more tightly conserved that amino acids 3-11, 22 and 23, suggesting that the latter may form a core recognition sequence that is of critical importance in the selection process. We do not know whether the other four residues participate directly in chain discrimination but this can be tested experimentally by site-directed mutagenesis.

The inventors have identified the functional domain which determines chain selectivity and show that trimerization is initiated via an interaction(s) between these

identified recognition sequences. It is unclear, however, whether the interactions which determine chain composition are the same as those which allow productive association and stabilization of the trimer. The nature of potential stabilizing interactions is uncertain, but recent data (Bulleid et al., 1996) indicate that, for type III procollagen at least, the formation of interchain disulfide bonds does not play a direct role in procollagen assembly. It has also been postulated that a cluster of four aromatic residues, which are conserved in the fibrillar collagens, collagens X, VIII and collagen like complement factor C1q, may be of strategic importance in trimerization.

The C-telopeptides were originally proposed to have a role in both procollagen assembly and in chain discrimination, the latter by virtue of the level of sequence diversity between various procollagen chains. However, the inventors have recently demonstrated (Bulleid et al., 1996) that the C-telopeptides of type III collagen do not interact prior to nucleation of the triple-helix, ruling out a role for this peptide sequence in the initial association of the C-propeptides. Data obtained from the assembly of hybrid chains indicates that the ability to discriminate between chains does not segregate with the species of C-telopeptide, lending support to this assertion.

Using this approach the inventors have been able to synthesize an entirely novel procollagen species compromising three proα2(I)Δ1 chains [proα2(I)Δ1]. Throughout this study procollagen 'mini-chains' with truncated triple-helical domains were used; however, the inventors have also demonstrated that full-length proα2(I) chains containing the 15-residue proα1(III) recognition sequence also self-associate into a triple-helical conformation (data not shown). Thus, the ability to introduce the chain recognition sequence into different pro-α chains provides the means to design novel collagen molecules with defined chain compositions. This, in turn, introduces the possibility of producing collagen matrices with defined biological properties, such as enhanced or differential cell-binding or adhesion properties. Furthermore, the identification of a short peptide sequence which directs the initial association between

procollagen chains may provide a target for therapeutic intervention allowing for the modulation or inhibition of collagen deposition.

The chimeric constructs described above may be used in the method of the present invention to allow the expression of exogenous procollagens in any cell-line without the problems associated with co-assembly with endogenously expressed procollagen. The uses of the methods of the invention are to express procollagen in cells either grown in culture or within tissues of the body. This will be of particular relevance for the production of recombinant procollagen in cell-lines such as fibroblasts which normally efficiently synthesis fibrillar collagens and in the treatment of collagen diseases by gene therapy.